The 1996 Speaker Recognition Evaluation Plan

Introduction

The emphasis this year will be on the issues of handset variation and test segment duration, as well as on the traditional evaluation goals, those being:

1. to drive the technology forward,
2. to measure the state-of-the-art, and
3. to find the most promising algorithmic approaches.

Technical Objective

The Task is to detect the presence of a hypothesized target speaker, given a segment of conversational speech over the telephone.

The research objective, given an overall 10% miss rate for target speakers, is to minimize the overall false alarm rate. The miss rate is a statistical random variable, however, not a control parameter. Therefore the performance is expressed instead in terms of a detection cost function (DCF), and the research objective is to minimize the DCF.

A secondary research objective is to achieve uniform performance across all target speakers.

The Evaluation

The task to be evaluated is the detection of a given target speaker. Given a test segment of speech, a target speaker identity will be assigned as a test hypothesis, and the task is to determine whether this test hypothesis is true or false.

Evaluation will be performed by measuring speaker detection performance in two related ways. Specifically, 1) detection performance will be measured directly from the correctness of target speaker detection decisions, and 2) detection performance will be characterized by an ROC:

- The correctness of the target speaker detection decisions will be measured by computing a detection cost function (DCF):

$$DCF = C_{Miss} \cdot P_{Miss|Target} \cdot P_{Target} + C_{FalseAlarm} \cdot P_{FalseAlarm|Non\text{-}Target} \cdot P_{Non\text{-}Target}$$

  where C_{Miss} and $C_{FalseAlarm}$ are cost factors. C_{Miss} will be 10 and $C_{FalseAlarm}$ will be 1 for this evaluation. DCF will be computed separately for each training condition. Also, a more detailed analysis will be performed by computing DCF separately for each test condition and each sex. The design goal and primary evaluation value for the target speaker *a priori* probability, P_{Target}, will be 0.01. In addition, NIST will compute values of DCF for a range of $\{P_{Target}\}$.

- An ROC will be constructed by pooling decision scores (for each of various training and test conditions) into two sets corresponding to whether the target speaker hypothesis was

true or false. These scores will then be sorted and plotted on PROC[1] plots. The conditions of interest include the 3 training conditions, the 3 test durations, the sex of the target speaker, whether or not the test handset was used in training, and whether or not the sex of the (non-target) test speaker is the same as that of the target speaker. PROCs will be compared by finding the minimum DCF and by measuring the false alarm probability at a miss probability of 0.10.

Please note that this task differs from previous speaker recognition tasks by requiring explicit decisions. Note also that scores from multiple target speakers are pooled **before** plotting PROCs. Thus it is critical that effective score normalization across speakers be achieved in order to achieve satisfactory speaker detection performance.

Evaluation Conditions

Training

There will be 3 training conditions for each target speaker. All 3 of these conditions will use 2 minutes of training speech data from the target speaker. The 3 conditions are:

- "**One-session**" training. Training data will be 2 minutes of speech data taken from only one conversation. This data will be stored in two files, with 1 minute of speech in each.
- "**One-handset**" training. Equal amounts of training data will be taken from two different conversations. These two conversations will use the same handset. (The same telephone number, actually). The training data for this *one-handset* condition will comprise the first of the *one-session* training files, plus an additional file containing 1 minute of speech from a different session (but from the same telephone number).
- "**Two-handset**" training. Equal amounts of training data will be taken from two different conversations collected using different handsets. (Different telephone numbers, actually.) The training data for this *two-handset* condition will comprise the first of the *one-session* training files, plus an additional file containing 1 minute of speech from a different session (and from a different telephone number).

The actual durations of the training files will vary from the nominal value of 1 minute, so that whole turns may be included whenever possible. Actual durations will be constrained to be within the range of 55-65 seconds.

Test

Performance will be computed and evaluated separately for female and male target speakers and for the 3 training conditions. For each of these training conditions, there are 3 different test conditions of interest. These are:

- Test segment duration. Performance will be computed separately for three different test durations. These durations will be nominally 3 seconds, 10 seconds, and 30 seconds. Actual duration will vary from nominal so that whole turns may be included whenever possible. Actual durations will be constrained to be within the ranges of 2-4 seconds, 7-13 seconds,

1. PROC plots are ROCs plotted on normal probability error (miss versus false alarm) plots.

and 25-35 seconds, respectively. A single turn will be used for the test segments whenever possible.

- Same/different handset. Performance will be computed separately for test segments which use the training handset versus those segments which use a different handset (as determined by the phone number that the speaker was using). The type of handset (same/different) being used in the test segment will be unknown to the system under test.
- Same/different sex. Performance will be computed separately for cross-sex versus same-sex non-target speakers. The sex of the speaker in the test segment will be unknown to the system under test.

Development Data

Development data is being supplied to help support R&D and to help prepare for the evaluation. This data may be used without restriction. It will be derived from the SwitchBoard Corpus. Both training and test segments will be supplied for about 45 male and 45 female speakers. These segments will be constructed by concatenating consecutive turns, as identified using the SwitchBoard transcription marks. Shorter duration segments will be taken from within the range of the corresponding longer duration segments. Each segment will be stored as a continuous speech signal in a separate SPHERE file. The speech data will be stored in 8-bit mu-law format. Auxiliary information will be included in the SPHERE headers to document the source file, start time and duration of all excerpts which were used to construct the segment.

The development data set will be supplied on a single CD-ROM, which will include both training and test data. The training data for each speaker will comprise four one-minute segments of speech, as described in the *Evaluation Conditions* section. There will be a total of about 200 test segments of each duration and for each sex.

Evaluation Data

The evaluation data will be of the same kind and in the same format as the development data. NIST will manually audit all segments to verify that the selected speech is for the identified speaker and does not include any significant extraneous speech from other speakers. There will be about 20 male and 20 female target speakers, and there will be about 200 male and 200 female non-target (impostor/background) speakers. All of these speakers will be different from the speakers in the development data set.

The evaluation corpus will be supplied on two CD-ROMs, by necessity. For convenience, male data will be stored on one disc and female data on the other. Knowledge of the sex of the target speaker is admissible side information and may be used if desired. However, such knowledge of the sex of the test segment's speaker is not allowed. (Non-target trials will include both same-sex and cross-sex data).

The evaluation data will include both training data and test data. Training data will be supplied for each of the target speakers, in the same form as that supplied for the development data set. Test data will be supplied for both the target and the non-target speakers, in the same form as that supplied for the development data set. There will be a total of about 1300 test

segments (about 800 segments from the target speakers and about 500 segments from the non-target speakers) for each sex and for each of the three test durations.

For all of the tests to be performed in this evaluation, there will be a grand total of about 20,000 target speaker trials and about 500,000 non-target speaker trials.

Evaluation Rules

A total of nine tests constitute the evaluation. These tests are namely a test for each of the three test durations for each of the three training conditions. Every evaluation participant is required to submit all of the results for each test.[1] In the event that a participating site does not submit a complete set of results, NIST will not report any results for that site.

The following evaluation rules and restrictions on system development must be observed by all participants:

- Each test segment is to be processed separately, independently, and without use of any knowledge of other test segments. Especially,
  - Normalization over multiple test segments is **not** allowed.
  - Normalization over multiple target speakers is **not** allowed.
- The use of transcripts for target speaker training is **not** allowed.
- Knowledge of the training condition **is** allowed.
- Side knowledge of the sex of the target speaker **is** allowed.
- Side knowledge of the sex or other characteristics of the test speaker is **not** allowed.
- Listening to the evaluation data, or any other experimental interaction with the data, is **not** allowed before all test results have been submitted. This applies to target speaker training data as well as test segments.
- None of the evaluation data supplied with this corpus, neither target-speaker data nor non-target-speaker data, may be used for non-target modelling.
- There is no limitation on the use of SwitchBoard data for the development speakers.
- There is no limitation on the use of data outside the SwitchBoard corpus and outside the SwitchBoard set of speakers.

Data Set Organization

Both the development data set (on one CD-ROM) and the evaluation data set (on two CD-ROM's) will have the same organization. Each disk's directory structure will organize the data

1. Funded contractors are required to do all nine of the tests. Unfunded participants are encouraged to do as many tests as possible. However, it is absolutely imperative that results for ***all*** of the target speakers and test segments in a test be submitted in order for that test to be considered valid and for the results to be accepted. If a participant anticipates being unable to complete all of the tests, NIST should be consulted for information about which tests to perform. Each participant must negotiate its test commitments with NIST before NIST ships the evaluation CDROM's to that site.

according to information admissible to the speaker recognition system. This directory structure will be as follows:

- There will be a single top-level directory on each disk, used as a unique label for the disk. This directory will be named **sid96d1** for the development data, **sid96e1m** for the male evaluation data, and **sid96e1f** for the female evaluation data.
- Under the top-level directory there will be two subdirectories, namely **train** (for storing the training data) and **test** (for storing the test data).
  - Under the **train** directory there will be two subdirectories, namely **male** (for the male speakers) and **female** (for the female speakers).
    - Under the **male** and the **female** directories there will be four subdirectories, namely **hs1_s1a** (for the first *one-session* training segment), **hs1_s1b** (for the second *one-session* training segment), **hs1_s2** (for the second *one-handset* training session), and **hs2** (for the segment from the second training handset).
      - In each of the **hs*** directories there will be one SPHERE-format data file for each of the speakers, containing approximately one minute of speech in mu-law format. The name of this file will be the ID of the speaker in the Switch-Board corpus, namely a simple 4-digit string, followed by "**.wav**".
  - Under the **test** directory there will be three subdirectories, namely "**30**" (for the 30 second test segments), "**10**" (for the 10 second test segments), and "**3**" (for the 3 second test segments).
    - In each of the **30**, **10**, and **3** segment duration directories there will be one SPHERE-format mu-law speech data file for each test segment. The names of these files will be pseudo-random alphanumeric strings, followed by "**.wav**".

For the development data set only, in each of the three **test** segment duration subdirectories there will be 6 index files for organizing and identifying the test segments. Each record in these files will contain the name of a test segment file and the ID for that file's speaker. These 6 index files will be named:

1. **male_hs1.ndx** (for male speakers using the first training handset)
2. **fema_hs1.ndx** (for female speakers using the first training handset)
3. **male_hs2.ndx** (for male speakers using the second training handset)
4. **fema_hs2.ndx** (for female speakers using the second training handset)
5. **male_hsx.ndx** (for male speakers using a non-training handset)[1]
6. **fema_hsx.ndx** (for female speakers using a non-training handset)[1]

For the evaluation data set only, in each of the three **test** segment duration subdirectories there will be 6 index files which specify the tests to be performed. Each record in these files will contain the name of a test segment file (in the corresponding test segment directory). These 6 index files will be named:

1. **male_1s.ndx** (for male targets using the *one-session* training condition)

1. Unfortunately, there are only very few development test segments (a handful at best) for speakers using handsets other than the two training handsets. Thus researchers will not have very helpful statistics for determining the proper speaker detection thresholds for the two-handset training condition. Consider this to be a challenge in "robust" decision making.

2. **fema_1s.ndx** (for female targets using the *one-session* training condition)
3. **male_1h.ndx** (for male targets using the *one-handset* training condition)
4. **fema_1h.ndx** (for female targets using the *one-handset* training condition)
5. **male_2h.ndx** (for male targets using the *two-handset* training condition)
6. **fema_2h.ndx** (for female targets using the *two-handset* training condition)

The evaluation test will be to process each test segment named in the index files against all of the target speakers of the indicated sex for the indicated training condition. For the case of same-sex trials, the same-sex index files will be an exhaustive list of all of the test segment files in the corresponding test segment directory. For the case of cross-sex trials, the cross-sex index files will contain a (possibly null) subset of test segment files.

Format for Submission of Results

Sites participating in the evaluation must report test results for all of the tests. These results must be provided to NIST in results files using a standard ASCII record format, with one record for each decision. Each record must document its decision with target identification, test segment identification, and decision information. Each record must thus contain seven fields separated by white space and in the following order:

1. The sex of the target speaker -- **M/F**
2. The training condition -- **1S/1H/2H** (*one-session*/*one-handset*/*two-handset*)
3. The target speaker ID (a 4-digit number)
4. The test segment duration -- **30/10/3**
5. The test segment file name
6. The decision (is the target speaker the same as the test segment speaker?) -- **T/F**
7. The score (where the more positive the score, the more likely the target speaker)

Execution Time

Sites must report CPU execution time for generating likelihood scores for the test data, as if the test were run on one CPU. Sites must also report the specs for the CPU as well as the memory, using a reporting format specified by NIST at the time of the evaluation.

Schedule

- The development data set CD-ROM will be distributed by NIST on 13 February 1996.
- Evaluation testing will begin on 27 February 1996, which is the date on which the evaluation data set CD-ROMs will be distributed by NIST.
- Evaluation testing will conclude by 11 March 1996, which is the deadline for submission of results to NIST. At this time NIST will distribute an answer key which associates a speaker ID with each test segment, to facilitate diagnostic analysis of the results.
- The follow-up workshop will be held on 27-28 March 1996 at the Maritime Institute.